The LIA RT'07 Speaker Diarization System

Corinne Fredouille[1] and Nicholas Evans[1,2]

[1] LIA-University of Avignon, BP1228, 84911 Avignon Cedex 9, France
[2] University of Wales Swansea, Singleton Park, Swansea, SA2 8PP, UK
(corinne.fredouille,nicholas.evans)@univ-avignon.fr

Abstract. This paper presents the LIA submission to the speaker diarization task of the 2007 NIST Rich Transcription (RT'07) evaluation campaign. We report a system optimised for conference meeting recordings and experiments on all three RT'07 subdomains and microphone conditions. Results show that, despite state-of-the-art performance for the single distant microphone (SDM) condition, in its current form the system is not effective in utilising the additional information that is available with the multiple distant microphone (MDM) condition. With post evaluation tuning we achieve a DER of 19% on the MDM task with conference meeting data. Some early experimental work highlights both the limitations and potential of utilising between-channel delay features for diarization.

1 Introduction

The speaker diarization task is an especially important contribution to the overall Rich Transcription (RT) paradigm, as evidenced by the RT evaluation campaigns administered by the National Institute of Standards and Technology (NIST). Also known as "Who spoke When", the speaker diarization task consists in detecting the speaker turns within an audio document (segmentation task) and in grouping together all the segments belonging to the same speaker (clustering task). Algorithms may be restricted to function with a single distant microphone (SDM) or have the potential to use multiple distant microphones (MDM).

Applied initially to conversational telephone speech and subsequently to broadcast news, the current focus is on conference and lecture meetings, tasks which pose a number of new challenges. Meeting room recordings often involve a greater degree of spontaneous speech with overlapped speech segments, speaker noise (laughs, whispers, coughs, etc.) and sometimes short speaker turns. Due to the availability of many different recording devices and room layouts, a large variability in signal quality has brought an additional level of complexity to the speaker diarization task and more generally to the RT domain.

This paper describes LIA's speaker diarization system and our experimental work and results relating to the most recent NIST RT'07 evaluation [1]. The system is developed on the conference meeting room datasets of the two previous RT campaigns and is evaluated on the three subdomains and three microphone conditions of the RT'07 datasets without modification. Results show that our system performs well on the single microphone condition (SDM), giving among the best results reported for the RT'07 evaluation. However, the system is shown not to be effective in utilising the additional information that is available with multiple microphones (the ADM and MDM conditions). Interestingly our system is shown to be relatively robust across the three subdomains including that of the coffee break condition newly introduced this year. Lastly, we present some post evaluation experiments which lead to improved diarization error rates across development and evaluation sets and some additional work that shows the limitations and potential of utilising the between-channel delay information.

The remainder of this paper is organized as follows: Section 2 presents the baseline E-HMM based speaker diarization system that was used for the RT'07 evaluation campaign. Section 3 describes the experimental protocol and presents results on the NIST RT'05 and RT'06 development datasets and the RT'07 evaluation dataset. Our post evaluation improvements are described in Section 4 and some initial experiments to assess potential of delay features in Section 5. Our conclusions are presented in Section 6.

2 Speaker diarization system

In this paper we report experimental work performed on the NIST RT'07 evaluation dataset. This includes three subdomains in addition to three different microphone conditions. These can involve single or multiple microphones. A number of strategies to utilise data from multiple microphones have appeared in the literature over recent years. The simplest involves the selection of a single channel according to some criteria, for example the channel with the highest estimated SNR. The channels may alternatively be combined according to an SNR-dependent weight parameter [2,3] and finally, with the provision for time delay of arrival (TDOA) estimation, channels may be resolved according to their respective delays before the addition, commonly referred to as the delay and sum beamforming approach [4]. In contrast to the work in [5] in which diarization was performed separately on each channel before fusing the outputs in a final post processing stage, all of these approaches aim to combine the multiple channels into a single channel prior to feature extraction and this seems to be the most dominant in the literature. For the experimental work reported here, where recordings from multiple microphones are available the different channels are processed very simply by summing related signals in order to yield a unique virtual channel which is used in all subsequent stages.

The LIA speaker diarization system was developed using the open source ALIZE speaker recognition toolkit [6]. The system is composed of 4 main steps:

- Speech/non-speech detection
- Pre-segmentation
- Speaker segmentation and clustering
- Post-normalisation and resegmentation

2.1 Speech/non-speech detection

The speech activity detection (SAD) algorithm employs feature vectors composed of 12 un-normalised Linear Frequency Cepstrum Coefficients (LFCCs) plus energy augmented by their first and second derivatives. It utilises an iterative process based on a Viterbi decoding and model adaptation applied to a two state HMM, where each state represents speech and non-speech events respectively. Each state of the HMM is initialised with a 32-component GMM model trained on separate data using an EM/ML algorithm. State transition probabilities are fixed to 0.5. Finally, some duration rules are applied in order to refine the speech/non-speech segmentation yielded by the iterative process.

2.2 Pre-segmentation

The pre-segmentation phase aims to provide an approximate speaker turn labelling to initialise and speed-up the subsequent segmentation and clustering stages. Now the signal is characterised by 20 LFCCs, computed every 10ms using a 20ms window. The cepstral features are augmented by energy but no feature normalisation is applied at this stage. A classical GLR criterion-based speaker turn detection is applied to two consecutive 0.5 second long windows with a 0.05 second step (single diagonal matrix Gaussian components). Relevant maximum peaks of the GLR curve are thus considered as speaker changes. Once speaker turns are detected, a local clustering process is applied in order to group together successive segments that are deemed to be sufficiently similar according to a thresholded GLR criterion.

2.3 Speaker segmentation and clustering

This step is the core of the LIA system. It relies on a one-step segmentation and clustering algorithm in the form of an evolutive hidden Markov model (E-HMM) [7, 8]: each E-HMM state aims to characterise a single speaker and the transitions represent the speaker turns.

This process, still based on 20 LFCCs plus energy coefficients, can be defined as follows:

1. Initialisation: The HMM has only one state, called L_0. A world model with 128 Gaussian components is trained on the entire audio show. The segmentation process is initialised with the segmentation outputs issued from the pre-segmentation stage and are utilised for the selection process.

2. Speaker addition: a minimum 3 second long candidate segment is selected among all the segments belonging to L_0 according to a likelihood maximisation criterion. The selected segment is attributed to L_x and is used to estimate the associated GMM model.

3. Adaptation/Decoding loop: The objective is to detect all segments belonging to the new speaker L_x. All speaker models are re-estimated through an adaptation process according to the current segmentation. A Viterbi decoding pass, involving the entire HMM, is performed in order to obtain a new segmentation. This adaptation/decoding loop is re-iterated while some significant changes are observed on the speaker segmentation between two successive iterations.

4. Speaker model validation and stop criterion: The current segmentation is analysed in order to decide if the new added speaker L_x is relevant, according to some heuristical rules on speaker L_x segment duration. The stop criterion is reached if there are no more minimum 3 second long candidate segments available in L_0 which may be used to add a new speaker; otherwise, the process goes back to step 2.

The segmentation stage is followed by a resegmentation process, which aims to refine the boundaries and to delete irrelevant speakers (e.g. speakers with too short speech segments). This stage is based on the third step of the segmentation process only: an HMM is generated from the segmentation and the iterative adaptation/decoding loop is launched. Here, an external world model, trained on microphone-recorded speech, is used for the speaker model adaptation. Compared to the segmentation process, the resegmentation stage does not utilise the pre-segmentation output. Indeed, all the boundaries (except speech/non-speech boundaries) and segment labels are re-examined during this process.

2.4 Post-normalisation and resegmentation

As reported in the literature [9], this last step consists in applying data normalisation drawing upon the speaker recognition domain. The resegmentation phase, described in the previous section, is repeated, but with a different parameterisation and now with data normalisation. Here the feature vector, comprising 16 LFCCs, energy, and their first derivatives, are normalised on a segment-by-segment basis to fit a zero-mean and unity-variance distribution. This segment-based normalisation relies on the output segmentation issued from the first resegmentation phase. The application of such a normalisation technique at the segmental level facilitates the estimation of the mean and variance on speaker-homogeneous data (compared with an estimate on the overall audio file involving many speakers).

3 System evaluation

This section presents the protocols and results for our submission to the NIST RT'07 evaluation campaign. Our development corpus is comprised of the conference meeting shows of the two previous, NIST RT'05 and RT'06 datasets. However, whilst the system is optimised only on conference meetings we have applied our system, without modification, to the three subdomains of the RT'07 evaluation, namely the conference and lecture meetings as in previous RT evaluations and, new to RT'07, data recorded during coffee breaks.

In its current form, our system is not capable of detecting overlapping speaker segments, thus our development work was optimised without scoring overlapping segments. However, since 2006 the primary metric of the RT evaluations includes the scoring of overlapping segments, thus scores are presented here with and without overlapping segments being taken into account. Unless otherwise stated, all scores referred to in the text are the scores *with* overlap taken into account. In addition, our recent research has been focused toward the conference meeting subdomain and the multiple distant microphone (MDM) condition thus we focus on this condition here.

3.1 Development Results

Table 1 illustrates diarization results for the NIST RT'05 and RT'06 datasets that were used for development. The second and third columns illustrate the missed and false alarm error as a percentage of scored speaker time and shows relatively stable performance across the two datasets with averages of 4.6% and 6.4% for the two datasets respectively (5.6% overall average) for missed speech errors and averages of 2.3% and 3.6% for false alarm errors (3.0% respectively).

There is, however, much greater variation in the results for speaker errors, as illustrated in the fourth column of Table 1. Across the two databases results range from 1.3% to 33.3%, though reasurringly the averages are relatively stable, at 13.3% and 11.6% for the RT'05 and RT'06 datasets respectively (12.4% overall average).

The final column illustrates the overall DER for each show and illustrates averages of 20.2% for RT'05 and 21.5% for RT'06 with an overall average of 20.9%. Thus across the two development datasets relatively consistent results are obtained.

3.2 Evaluation Results

Turning to the RT'07 evaluation we observe similar levels of performance for both missed and false alarm errors with values of 4.5% and 2.0% respectively. However, there is a significant increase in the speaker error which rises from a 12.4% average across the two development sets to 17.7% for the RT'07 evaluation set. Once again there is a high level of variation between the best and worst results which now range from 3.7% to 36.9%. The degradation in speaker error when moving from the development to the evaluation sets accounts for an increase in the overall DER from 20.9% across the two development sets to 24.2% for the RT'07 evaluation set.

Table 3 summarises the results obtained by our system on each of the three subdomains of the RT'07 evaluation for each microphone condition. Only the primary metric which includes overlapped segments is given. The first general observation is that there is very little difference in performance between the different microphone conditions. This indicates that the system is not effective in utilising the additional information that is available in the additional channels. For the conference meeting subdomain a performance of 24.2% with the MDM condition compares to 24.5% with the SDM condition, an insignificant difference. In addition, the missed, false alarm and speaker error rates are close and, with the

Table 1. Missed speaker, false alarm and speaker error rates for the RT’05 and RT’06 datasets as used for development. Also, overall average time weighted across the two datasets. Results with/without scoring overlapping segments.

Show	Missed	FAlarm	Speaker	Overall
RT’05				
AMI_20041210	1.0/0.6	0.9/0.9	1.3/1.3	3.2/2.8
AMI_20050204	3.4/1.3	0.9/1.0	33.3/34.6	37.7/36.9
CMU_20050228	11.1/5.2	0.9/1.0	5.7/6.2	17.7/12.5
CMU_20050301	3.3/0.6	1.8/1.9	13.0/13.8	18.1/16.3
ICSI_20010531	6.3/4.3	3.0/3.2	13.0/13.5	22.4/20.9
ICSI_20011113	8.0/1.1	2.5/2.9	29.1/32.3	39.6/36.4
NIST_20050412	6.8/0.0	3.8/4.4	1.9/2.1	12.4/6.5
NIST_20050427	2.9/0.3	6.1/6.5	6.9/7.3	15.9/14.2
VT_20050304	0.7/0.4	1.1/1.2	8.9/8.9	10.7/10.5
VT_20050318	3.2/2.5	2.2/2.3	25.8/26.0	31.2/30.8
RT’05 average	**4.6/1.6**	**2.3/2.5**	**13.3/14.0**	**20.2/18.0**
RT’06				
CMU_20050912	11.1/0.1	6.4/8.1	10.0/11.3	27.5/19.5
CMU_20050914	9.8/0.7	3.0/3.6	4.3/4.2	17.1/8.4
EDI_20050216	5.0/1.6	1.5/1.6	21.6/22.6	28.1/25.7
EDI_20050218	4.4/1.0	2.5/2.7	10.7/10.7	17.6/14.5
NIST_20051024	6.6/0.5	1.7/2.0	8.7/9.3	17.0/11.8
NIST_20051102	5.1/0.2	3.5/3.9	21.3/22.9	29.9/26.9
VT_20050623	4.6/0.4	7.4/8.0	3.5/3.3	15.5/11.7
VT_20051027	3.2/1.5	2.9/3.0	11.0/11.0	17.13/15.5
RT’06 average	**6.4/0.7**	**3.6/4.0**	**11.6/12.2**	**21.5/17.0**
Overall Average	**5.6/1.2**	**3.0/3.3**	**12.4/13.1**	**20.9/17.5**

Table 2. Missed speaker, false alarm and speaker error rates for the RT’07 evaluation as submitted. Results with/without scoring overlapping segments.

Show	Missed	FAlarm	Speaker	Overall
RT’07				
CMU_20061115-1030	7.4/0.2	4.6/5.4	9.7/9.8	21.8/15.4
CMU_20061115-1530	3.3/0.0	5.1/5.5	14.5/15.0	23.0/20.6
EDI_20061113-1500	8.9/2.0	0.8/0.9	22.8/25.0	32.5/27.9
EDI_20061114-1500	3.2/1.1	1.8/1.9	23.3/23.9	28.4/26.9
NIST_20051104-1515	3.8/0.6	0.9/0.9	7.6/8.0	12.2/9.5
NIST_20060216-1347	2.5/0.7	1.4/1.5	20.9/21.6	24.8/23.8
VT_20050408-1500	1.5/1.1	0.6/0.6	36.9/37.1	39.0/38.8
VT_20050425-1000	5.5/1.0	0.7/0.8	3.7/3.9	9.9/5.6
RT’07 average	**4.5/0.8**	**2.0/2.2**	**17.7/18.6**	**24.2/21.5**

Table 3. Summary of performance for the three conditions as submitted to the NIST RT'07 evaluation.

Subdomain	Mic. Cond.	Missed	FAlarm	Speaker	Overall
Conference meeting	MDM	4.5	2.0	17.7	24.2
	SDM	4.7	2.1	17.7	24.5
Lecture meeting	ADM	4.1	7.2	19.3	30.5
	MDM	3.4	6.9	20.9	31.2
	SDM	3.6	6.5	19.4	29.5
Coffee break	ADM	3.5	3.6	19.2	26.4
	MDM	3.0	5.0	17.5	25.5
	SDM	3.3	4.6	18.4	26.3

exception of the false alarms for the lecture meeting condition, this observation is consistent across the three subdomains.

Whilst the best overall performance is obtained with conference meeting data (the same subdomain on which the system was developed), similar levels of performance are observed with coffee break data with only a marginal decrease in performance to 25.5% for the MDM condition. However, for lecture meeting data there is a marked degradation in performance to 31.2% for the same condition. This is attributed predominantly to an increase in the false alarm error rate whilst the missed and speaker error rates remain relatively consistant. An increase in the false alarm error rate could be expected due to increased levels of activity and noise for this condition.

4 Post Evaluation Improvements

According to the results of the system on the RT'07 corpus, two aspects of the speaker diarization system have been studied. The first is related to the successive segment clustering of the pre-segmentation step. As reported in Section 2.2, the pre-segmentation phase is applied in order to speed up the segmentation and resegmentation steps, which is the core of the speaker diarization system. It involves both a speaker turn detection and a coarse clustering, applied locally to aggregate successive segments. This clustering is based on a thresholded GLR criterion and provides a segmentation output, which is involved in the subsequent segmentation phase. The quality of this segmentation output is relevant for the later process, and is constrained by the threshold value used for the clustering.

The second aspect on which we focus, is the selection strategy involved in the segmentation process when adding a new speaker to the E-HMM. This selection strategy is still an issue, since it can contribute largely to the overall system performance. Indeed, the selection of an irrelevant segment (for instance, the selection of a non-speech segment, mislabeled by the speech/non-speech detection or of a multi-speaker segment, due to clustering misclassification) may dramatically disturb the segmentation process. Compared with previous versions of the speaker diarization system [3], hypothesized segments for the selection process may be of variable durations (with a minimum fixed to 3 seconds), since they are directly issued from the pre-segmentation step[3].

Both these aspects are strongly correlated since the clustering process will provide the selection strategy with hypothesized segments. According to the clustering threshold, the number of segments may vary as well as their quality (in terms of speaker purity). These two factors are very important for the selection but also for the overall segmentation process. Indeed, the number of segments available for the selection indirectly determines the number of speakers, which may be potentially added to the E-HMM. In the same way, the less pure the segments, the less robust the speaker models.

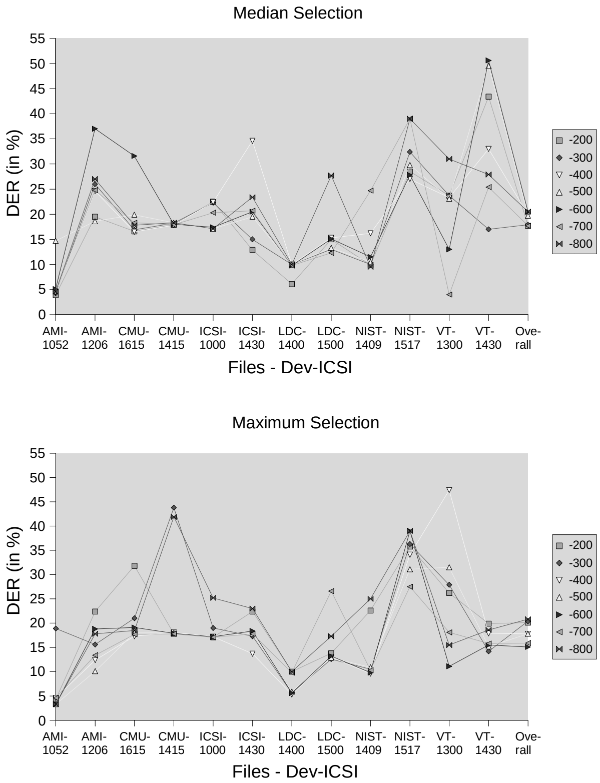
In this section, we compare performance of the speaker diarization system according to various values for the clustering threshold and two different selection strategies. For the latter, the maximum likelihood criterion, named "Maximum Selection", used for the evaluation campaign as reported in Section 2.3, is compared with an averaged likelihood criterion, named "Median Selection". In the last case, the segment which is close in terms of likelihood to the likelihood mean computed over all the hypothesized segments is selected.

These post-evaluation experiments have been conducted on the RT'07 evaluation data set as well as on the development set proposed by ICSI in [10]. The latter has been chosen in order to be able to compare the performance of the speaker diarization system with that of ICSI's system (named $DevICSI$ in the rest of the paper), regarding the performance gap drawn by the evaluation data set.

Figures 1 and 2 provide the speaker diarization performance in terms of DER involving overlapping segments and according to different configurations (threshold values and selection strategies). Different remarks can be pointed out from these figures:

- threshold values may largely influence scores, depending on the files. Regarding *VT-1500/Dev-ICSI*, the DER varies from 2.9% to 38%. On the opposite, some files (e.g. *NIST-1515/RT07*, *CMU-1530/RT07*, or *LDC-1400/Dev-ICSI* are less sensitive to threshold variation, leading to quite stable DER scores whatever configuration used ;
- for a given clustering threshold, selection strategies may behave in opposite manner. For instance, given value -600 for the threshold, the median selection gets 50.6% DER on *VT-1430* files against 15.5% DER for the maximum selection;
- optimal (from an empirical point of view) threshold is different according to the selection strategies and data sets observed (-600 for the Maximum selection against -200 or -700 for the Median selection regarding *Dev-ICSI* data set and -300 for the maximum selection and -200 for the median one

[3] In the previous versions of the speaker diarization system, the hypothesized segments were extracted directly from the speech portions of signal issued from the speech/non-speech detection process. Their length was fixed to 3seconds.


Median Selection

Maximum Selection

Fig. 1. Speaker diarization performance (DER in %) on *Dev-ICSI* according to selection strategies (Maximum or median selection) and clustering thresholds.

for the RT'07 data set). In this way, the official LIA score on the RT'07 is enhanced from 24% to 19.2% DER;

The pre-segmentation responds to its initial goal, which was to speed-up the segmentation step by diminishing hypothesized segments for the selection phase and expanding their length. Compared with previous versions of the speaker diarization system, which will test all the 3second segments available, CPU time was decreased from 3*RT to 0.25*RT (with the clustering threshold fixed to -200). Moreover, speaker diarization tests performed on RT'05 and RT'06 (not reported here) show no loss of performance in this case. Nevertheless, experiments reported in this section outlines an unstable behavior of the speaker diarization system, depending on the clustering threshold value, on the data sets observed or on the selection strategy chosen for the process. Further investigation will focus on novel solutions for pre-segmentation enhancement, which still respond to the speed request while preserving speaker diarization behavior stability.

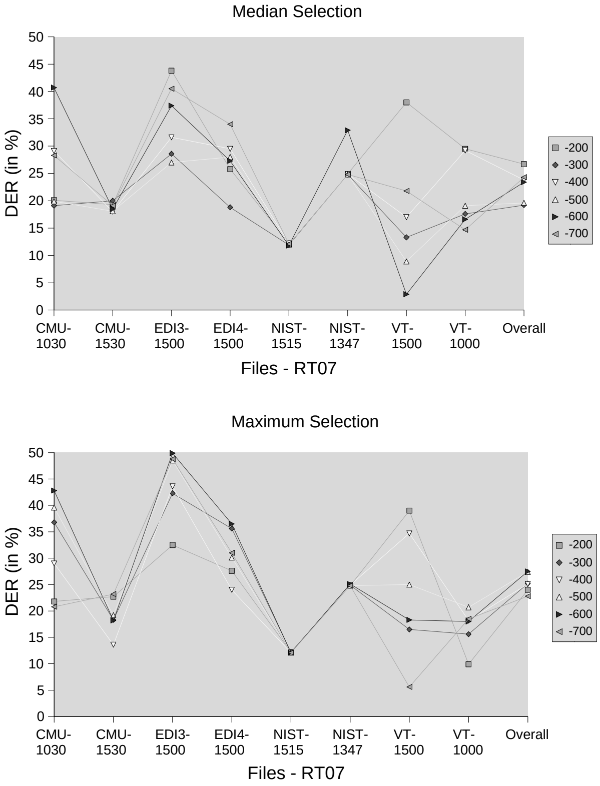


Fig. 2. Speaker diarization performance (DER in %) on RT'07 according to selection strategies (Maximum or median selection) and clustering thresholds.

5 Between-channel delay features

Estimates of the between-channel delay characterise each speaker's position in the room and thus may be utilised as features to assist diarization and in this section we report LIA's early work to assess the potential. The use of delay features has been reported before, for example in [11] and more recently in [10,12] in which a DER of 31% was reported on the NIST RT'05s conference room evaluation dataset. These results do not compare well to a DER of 18% with acoustic-only features following delay and sum beamforming but nonetheless offer additional information which, when used with the acoustic features in a combined log-likelihood, leads to an improved DER of 15%. Thus whilst the delay features produce only a small improvement in DER this work clearly highlights the potential.

Here we report some initial experiments with our diarization system using delay features. This is very much embryonic work and, in moving from acoustic to delay features, we have not modified the underlying diarization system in any way other than to handle feature vectors of varying order. In all cases delay features are estimated using the conventional generalised cross correlation phase transform (GCC-PHAT) approach [13]. Four different experiments are reported.

We first seek to evaluate the potential of delay features in a 'fake' experiment using the key to identify each speakers' segments, to estimate the delay characteristics of each speaker and then to perform diarization without using the speaker labels but using the segment boundaries identified by the key. Delay features are estimated using whole segments and speaker models are derived from the median between-channel delays. The classifier is based simply on the minimum Euclidean distance between segments and speaker models. The GCC-PHAT algorithm requires a reference channel and results are shown in Table 4 for where the SDM channel is used as a reference (column 2) and where a reference channel is selected automatically (column 3). Here the reference is the channel which exhibits the highest correlation to all other channels. Results show that there is little difference between the two sets of results, each producing a diarization error rate of 15%. The second observation relates to the large variation in the performance, with a best performance of 2% and particularly poor performance being achieved with the two EDI and final VT shows. This leads to the conclusion that in these three shows a number of speakers are difficult to separate in delay space. Given that in these two experiments we have used the key to identify speaker segments these results serve to highlighting the potential limitation and difficulty of using delay features.

The fourth column in Table 4 illustrates the results of a 'real' experiment where now the delay features come from sliding frames of 200 ms in length and with a rate of 100 frames per second. Median filtering is used to smooth the delay profiles. Delay features are calculated with a reference channel that is automatically selected via correlation. As would be predicted the results are much worse and show DERs of between 28% and 56% with an average of 41%. This result does not compare favourably with that of the acoustic-only features (DER of 19%) presented in Section 4 and in the final set of experiments we seek to evaluate the potential of combining the acoutic and delay features.

Using weights of 0.9 for the acoustic and 0.1 for the delay features the two streams are combined in the segmentation and resegmentation stages with a joint log likelihood as in [10]. DER results are presented in the final column of Table 4 which show an overall average DER of 31%. However, without having modified our system in any way other than to fascilitate the combination, this result is hardly surprising. It clearly illustrates the difficulty in accurately estimating and making appropriate use of the delay due to the different nature of acoustic and delay features which are likely to require fundamentally different approaches to handle. This area is a topic of future work.

6 Conclusions

This paper presents the results of LIA's speaker diarization system on the NIST RT'07 evaluation dataset. The system is shown to give state-of-the-art perfor-

Table 4. Diarization performance on the RT'06 database using delay features with SDM channel as reference and segments identified using the key (column 2), the same except with an automatically chosen reference channel (column 3), a real experiment with delay features only (column 4) and for combined acoustic and delay features (column 5).

Show	SDM ref fake	Auto ref fake	D real	A+D real
RT'06				
CMU_20050912	7.8	7.8	55.6	33.8
CMU_20050914	3.3	3.3	28.0	22.1
EDI_20050216	25.0	25.0	48.1	26.0
EDI_20050218	43.7	43.7	50.4	17.6
NIST_20051024	10.8	2.2	24.4	46.3
NIST_20051102	2.3	4.8	42.7	46.6
VT_20050623	6.5	13.7	43.5	15.3
VT_20051027	22.7	22.7	33.7	29.6
Overall average	**15.3**	**15.2**	**40.8**	**30.5**

mance on the single distant microphone condition but that it is not effective in making use of the additional information provided by multiple channels. An overall average diarization error rate of 24% is reported on the multiple distant microphone condition for conference meeting data. With post evaluation tuning this figure falls to 19% with consistant results obtained across development data. In addition, consistant results are reported across the three subdomain tasks with only a negligible degradation in results observed with lecture meeting data. Finally some of LIA's early experiments with delay features are reported which illustrate the limitations and potential of utilising between-channel delay features for diarization. Our future work is focused toward properly harnessing the additional information in multiple channels in order to improve the stability of the system and hence fully realise the potential of a system proven to give state-of-the-art results for a single microphone channel.

References

[1] NIST: 2007 (RT'07) Rich Transcription meeting recognition evaluation plan. http://www.nist.gov/speech/tests/rt/rt2007/docs/rt07-meeting-eval-plan-v2.pdf (2007)

[2] Istrate, D., Fredouille, C., Meignier, S., Besacier, L., Bonastre, J.F.: NIST RT'05S evaluation: pre-processing techniques and speaker diarization on multiple microphone meetings. In: Machine Learning for Multimodal Interaction: Second International Workshop, Springer-Verlag, Edinburgh, UK. (2005)

[3] Fredouille, C., Senay, G.: Technical improvements of the e-hmm based speaker diarization system for meeting records. In: MLMI'06, Washington, USA (2006)

[4] Anguera, X., Wooters, C., Hernando, J.: Proc. ASRU'05. In: Speaker diarization for multi-party meetings using acoustic fusion. (2005)

[5] Fredouille, C., Moraru, D., Meignier, S., Besacier, L., Bonastre, J.F.: The NIST 2004 spring rich transcription evaluation : two-axis merging strategy in the context of multiple distance microphone based meeting speaker segmentation. In: RT2004 Spring Meeting Recognition Workshop. (2004) 5

[6] Bonastre, J.F., Wils, F., Meignier, S.: ALIZE, a free toolkit for speaker recognition. In: ICASSP'05, Philadelphia, USA (2005)

[7] Meignier, S., Bonastre, J.F., Fredouille, C., Merlin, T.: Evolutive HMM for speaker tracking system. In: ICASSP'00, Istanbul, Turkey (2000)

[8] Meignier, S., Moraru, D., Fredouille, C., Bonastre, J.F., Besacier, L.: Step-by-step and integrated approaches in broadcast news speaker diarization. Special issue of Computer and Speech Language Journal, Vol. 20-(2-3) (2006)

[9] Zhu, X., Barras, C., Meignier, S., Gauvain, J.L.: Combining speaker identification and BIC for speaker diarization. In: EuroSpeech'05, Lisboa, Portugal (2005)

[10] Pardo, J.M., Anguera, X., Wooters, C.: Proc. ICSLP'06. In: Speaker diarization for multiple distant microphone meetings: mixing acoustic features and inter-channel time differences. (2006)

[11] Ellis, D.P.W., Liu, J.C.: Speaker turn detection based on between-channel differences. In: Proc. ICASSP'04. (2004)

[12] Pardo, J.M., Anguera, X., Wooters, C.: Proc. MLMI'06. In: Speaker diarization for multi-microhpone meetings using only between-channel differences. (2006)

[13] Brandstein, M.S., Silverman, H.F.: Proc. ICASSP'97. In: A robust method for speech signal time-delay estimation in reverberent rooms. (1997)